

An automatic adaptive method to combine summary statistics in approximate Bayesian computation

Jonathan U. Harrison · Ruth E. Baker

Received: date / Accepted: date

Abstract To infer the parameters of mechanistic models with intractable likelihoods, techniques such as approximate Bayesian computation (ABC) are increasingly being adopted. One of the main disadvantages of ABC in practical situations, however, is that parameter inference must generally rely on summary statistics of the data. This is particularly the case for problems involving high-dimensional data, such as biological imaging experiments. However, some summary statistics contain more information about parameters of interest than others, and it is not always clear how to weight their contributions within the ABC framework. We address this problem by developing an automatic, adaptive algorithm that chooses weights for each summary statistic. Our algorithm aims to maximize the distance between the prior and the approximate posterior by automatically adapting the weights within the ABC distance function. To demonstrate the effectiveness of our algorithm, we apply it to several stochastic models of biochemical reaction networks, and a spatial model of diffusion, and compare our results with existing algorithms.

Keywords Approximate Bayesian Computation · Summary Statistics · Sequential Monte Carlo · Likelihood Free

1 Introduction

When using quantitative models to explore biological or physical phenomena, it is crucial to be able to estimate parameters of these models and account appropriately for uncertainty in both the parameters and model predictions. Bayesian statistics offers a wealth of tools in this regard [15, 31]. Bayes' theorem gives us that the posterior, $p(\theta|D)$, of parameters, θ , given data, D , is proportional to a prior, $\pi(\theta)$, on the parameters multiplied by the likelihood, $p(D|\theta)$, of data, D , given those parameters: $p(\theta|D) \propto p(D|\theta)p(\theta)$. The prior represents our beliefs about the parameters prior to observing the data, the likelihood gives the probability of observing the data, given a certain set of parameters, and these result in the posterior, which returns updated beliefs about the parameters after having observed the data.

However, much of the current theory surrounding the generation of posterior distributions for parameter inference relies on being able to evaluate the likelihood of the data given the parameters of a model. In practice, for a large class of mechanistic models the likelihood is not tractable, either due to computational or analytical complexity. Therefore, the use of likelihood-free methods for inference, including approximate Bayesian computation (ABC) [5, 6, 26, 28, 30], indirect inference [13], synthetic likelihoods [25, 32], particle Markov Chain Monte Carlo (pMCMC) [1, 2, 12, 23], expectation propagation [4], and other similar methods, has become widespread [14]. In particular, ABC has been widely adopted due to its ease of understanding and implementation.

Jonathan U. Harrison
Mathematical Institute,
University of Oxford
E-mail: harrison@maths.ox.ac.uk

Ruth E. Baker
Mathematical Institute,
University of Oxford

1.1 Approximate Bayesian computation

Suppose we wish to infer a posterior distribution over parameters θ of a generative model such that we can simulate from $\mathbf{x} \sim f(\mathbf{x}|\theta)$. In ABC, parameters θ are drawn from a prior, $\pi(\theta)$, and data, x^* , is simulated from the generative model using those parameters, such that $x^* \sim f(x|\theta)$. The distance between the simulated dataset, x^* , and the real data, y , is calculated using a distance function $d(x^*, y)$. If this distance is less than a certain tolerance, ϵ , then the parameters θ can be accepted into the approximate posterior sample. Choice of the tolerance ϵ can be avoided, to some extent, by simulating a fixed, large number, N , of parameter samples and datasets, calculating the corresponding distances for these and accepting the proportion α of those that lie closest to the real data. We will use this approach here.

Algorithm 1 summarizes how samples from an approximate posterior can be generated via a more efficient version of ABC using sequential Monte Carlo techniques, known as ABC-SMC [9, 27, 29]. Importance sampling is used iteratively so that instead of sampling repeatedly from the prior, parameters are sampled from an approximate posterior at each generation of the algorithm. A weight must be given to each sample to correct for the fact that it is not drawn from the prior.

Algorithm 1 ABC-SMC

- 1: Set generation index $t = 0$. Set a proportion of samples to accept, α .
- 2: Set particle index $i = 1$.
- 3: If $t = 0$, sample $\theta^{**} \sim \pi(\theta)$. Proceed to step 4.
Else sample θ^* from previous population θ_{t-1}^i with weights v_{t-1} .
Perturb θ^* to give $\theta^{**} \sim K_t(\theta|\theta^*)$. If $\pi(\theta^{**}) = 0$, return to step 3.
- 4: Simulate dataset $x^{*i} \sim f(x|\theta^{**})$ and calculate distance $d(y, x^{*i})$.
- 5: Set $\theta_t^i = \theta^{**}$.
Calculate the weight v_t^i for particle θ_t^i via:

$$v_t^i = \begin{cases} 1, & \text{if } t = 0, \\ \frac{\pi(\theta_t^i)}{\sum_{j=1}^N v_{t-1}^j K_t(\theta_t^j, \theta_t^i)}, & \text{if } t > 0. \end{cases}$$

- If $i < N$, set $i = i + 1$ and return to step 3.
- 6: Select the proportion α of samples closest to the real data to keep and reject the rest, resulting in $M = \alpha N$ samples.
 - 7: Normalize the particle weights v_t^i .
If $t < T$, then set $t = t + 1$. Go to step 2.
Else stop.
-

1.2 Choice of summary statistics

Suppose we are interested in inferring multi-dimensional parameters for a model that we can simulate, but cannot evaluate the likelihood directly. In many practical circumstances, the data (either collected experimentally or simulated from the *in silico* model) will be very high dimensional. High-dimensional data poses difficulties within the ABC framework, as it is difficult to sensibly estimate when the output of a particular simulation is ‘close’ to the data. This issue is further compounded using stochastic models. As such, it is often necessary to work with lower-dimensional summary statistics of data. Examples of these summary statistics may be data points within a time series, an average transition time between different states of a system, or the moments of a certain species within a model. However, not all summary statistics are equally informative about the posterior, which we aim to infer. Common practice is to combine summary statistics based on some heuristic approach, such as the weighting the contribution of each summary statistic according to its standard deviation. However, it is not clear whether these heuristic approaches result in optimal weighting of the various summary statistics available. As such, the aim of this work is to provide an automated and adaptive method for determining the weighting of available summary statistics in order to optimize the quality of the resulting posterior.

Previous work has also considered how to weight summary statistics for ABC. For example, a genetic algorithm has been used to choose the weights of different summary statistics [18]. This genetic algorithm attempts to optimize the mean squared error (MSE) of the posterior samples from the true parameter, however this may not be known in practice. A method for adaptively choosing summary statistic weights for ABC based on the scale of the summary statistics has also been investigated [24]. The median absolute deviation, a measure of spread of a statistic, is used for the scaling. The aim is that all summary statistics contribute equally to the distance function, but in practice, this may not be the most desirable choice, as some summary statistics are clearly more informative than others.

1.3 Outline

Our contribution in this work is to present a flexible, novel framework for improving inference with ABC by adapting the weights of different summary statistics to maximize the gain in posterior information from a dataset. This is helpful for avoiding bias and variance from redundant information in data (such as would be

the case when including a summary statistic that is uncorrelated with the parameters of interest). A further advantage of our work is that it can alleviate the burden of designing and selecting summary statistics ‘by hand’, since a large collection of summaries can be used and weighted appropriately via our procedure. It is also possible to combine our framework with existing dimensionality reduction techniques for summary statistics in ABC (see Section 4.1).

We outline in Section 2 our adaptive algorithm for combining summary statistics in an ABC framework. To demonstrate the utility of our algorithm, we apply it to several test problems based on biochemical reaction networks in Section 3. We compare results of parameter inference using our algorithm against benchmark results from applying ABC-SMC using other choices of weights for the summary statistics. Finally, in Section 4, we summarize the work presented in this article and compare our methodology for combining summary statistics with other techniques in the literature that are based on dimensionality reduction of a set of summary statistics.

2 An algorithm for automatic weighting of summary statistics

In order to use ABC-SMC (see Algorithm 1), we must specify a function to measure the distance between simulated and real datasets. Suppose we take a weighted Euclidean distance as our ABC distance function such that $d(x_1, x_2) = \sum_i w_i (x_{1i} - x_{2i})^2$, where the sum over i is taken over all the summary statistics considered¹. This is a reasonable and flexible choice commonly used in the literature [20]. It is these distance weights, w_i , that control how the summary statistics are combined in this case. Given simulated pairs of parameter samples and datasets, we perform a search through the space of possible weights, w_i , to find weights that maximize a distance between the prior and the posterior that represents the maximum possible gain in information about the parameters from the given data. Constructing the weights in this way allows us to account for the scale of the summary statistics, as well as their relative contribution to a posterior.

2.1 Adaption of weights

We seek to optimize the weights, w_i , so that we can place less emphasis on summary statistics that are not

¹ We will treat a vector of summary statistics \mathbf{x} as consisting of scalar values $\mathbf{x} = (x_1, \dots, x_K)$.

informative for the posterior, but also scale summary statistics appropriately so that we do not neglect information about certain parameters. We do this within the ABC-SMC framework [9, 27, 29] given in Algorithm 1. We outline our proposed methodology in Algorithm 2.

At each generation, we search for the weights, w_i , of the distance function that maximize the distance between the prior and resulting posterior, given N ABC samples from the model for different θ values. This distance between prior and posterior gives a measure of the information gain in moving from the prior to the posterior. We specify an objective, L , to maximize, where $L = d - \lambda|\mathbf{w}|$, such that d is the distance between prior samples and approximate posterior samples, and λ is a scalar parameter enforcing regularization of the weights. Regularising the weights should force the weights to be sparse, so that if a summary statistic is uncorrelated with the parameters, the corresponding weight should be close to zero [7]. The heuristic used to perform the search through weight space in our implementation of Algorithm 2 is a random walk proposal, where proposed weights are accepted if they improve the objective.

We use the Hellinger distance to measure the discrepancy between the prior and posterior, and so to detect the optimality of our posterior. The Hellinger distance is defined, for distributions P and Q , with densities f and g , respectively, as

$$H^2(P, Q) = \frac{1}{2} \int \left(\sqrt{f(x)} - \sqrt{g(x)} \right)^2 dx, \quad (2)$$

and can be computed for discrete probability vectors P and Q as

$$H(P, Q) = \frac{1}{\sqrt{2}} \|\sqrt{P} - \sqrt{Q}\|_2.$$

Alternative measures between distributions such as the Euclidean distance or Kullback-Leibler (or KL) divergence can be used. In our experience, the Hellinger distance performs better than alternatives, particularly for robustly identifying relatively small differences between posterior distributions when weights are optimized, an observation that is supported by other work [16].

3 Examples

We apply our algorithm of automatic, adaptive weighting of summary statistics to a variety of test problems based on different chemical reaction networks. The dynamics of these networks are simulated stochastically using Gillespie’s direct method [11], which allows us to sample trajectories directly from the model. Although for some of these models it is possible to solve for the

Algorithm 2 Adaption of distance weights for ABC

- 1: Set generation index $t = 0$. Set a proportion of samples to accept, α .
- 2: Set particle index $i = 1$.
- 3: If $t = 0$, sample $\theta^{**} \sim \pi(\theta)$. Proceed to step 4.
Else sample θ^* from previous population θ_{t-1}^i with weights v_{t-1}^i .
Perturb θ^* to give $\theta^{**} \sim K_t(\theta|\theta^*)$. If $\pi(\theta^{**}) = 0$, return to step 3.
- 4: Simulate dataset $x^{*i} \sim f(x|\theta^{**})$.
- 5: Set $\theta_t^i = \theta^{**}$.
Calculate weight v_t^i for particle θ_t^i via:

$$v_t^i = \begin{cases} 1, & \text{if } t = 0, \\ \frac{\pi(\theta_t^i)}{\sum_{j=1}^N v_{t-1}^j K_t(\theta_t^j, \theta_t^i)}, & \text{if } t > 0. \end{cases} \quad (1)$$

- If $i < N$, set $i = i + 1$ and return to step 3.
- 6: Initialize search index $k = 1$ and distance weights $w_k^j = 1, \forall j$.
- 7: Search through weight space by perturbing the weights using $\mathbf{w}_{k+1} \sim \tilde{K}_k(\mathbf{w}|\mathbf{w}_k)$ and set $k = k + 1$.
- 8: For given distance weights \mathbf{w}_k , calculate the distance between simulated data and real data for each pseudo dataset, $d(y, x^{i*})$.
- 9: Select the proportion α of the samples closest to the real data to keep and reject the rest, resulting in $M = \alpha N$ samples (given distance weights \mathbf{w}_k).
- 10: Evaluate the distance, d_k , between the approximate posterior at generation t and the prior.
Record the objective to maximize,

$$L_k = d_k - \lambda|\mathbf{w}_k|,$$

where λ is a regularization parameter.

- 11: If $k < K$, then return to step 7. Else select the index \hat{k} which maximizes the objective L and select the corresponding weights $\mathbf{w}_{\hat{k}}$.
Calculate the corresponding parameter samples, $\theta_{\hat{k}}^i$, and particle weights, $v_{\hat{k}}^i$, using (1).
- 12: Normalize the particle weights $v_{\hat{k}}^i$.
If $t < T$, then set $t = t + 1$. Go to step 2.
Else stop.

likelihood analytically, we attempt parameter inference by simulation, since solving for the likelihood is very computationally expensive and, in general, the analytical solution is not available. The summary statistics collected for each problem are in the form of a time series, to imitate data that could be collected from a biological experiment.

To demonstrate the effectiveness of taking a flexible choice of distance weights, we make two comparisons. Firstly, we compare results obtained using Algorithm 2 to those generated using a uniform choice of weights: $w_i = 1 \forall i$. Secondly, we compare to results generated using weights that scale with each summary statistic. Here we use $w_i = 1/\sigma_i \forall i$, where σ_i is the standard deviation of the given summary statistic. This is a frequently used choice of weight for summary statistics [6].

We note that a table summarising the parameters used in the implementation of all the test problems can be found in Appendix A.

3.1 Death process

For our first test problem, we consider estimating the rate parameter for a single, first order degradation reaction:



We will consider for this, and subsequent, test problems that time has been non-dimensionalized. Initially, we assume there are $A(0) = 10$ particles in the system, which is observed over a (non-dimensional) time period $[0, 20]$. We assume it is possible to measure the state of the system (in this case the number of molecules of species A) without observation noise at given time points t_0, t_1, \dots, t_n . For this test problem, we assume that we measure at n equally spaced time intervals, and take $n = 32$.

As our summary statistics, we take $S = [A(t_0), A(t_1), \dots, A(t_n), z]$ where z is an observation of a random variable $Z \sim N(0, \sigma^2)$ that is uncorrelated with the death process. We suppose that the scale of the variance, σ , is different to the scale of the observations of the exponential decay process, giving a simple system with a two-dimensional parameter to infer: $\theta = (k, \sigma)$. Note that the scale of z is determined by the standard deviation σ , but the scale of death process is affected by the initial condition $A(0) = 10$, resulting in two distinct scales in these summary statistics.

Results of parameter inference for this system using ABC-SMC are shown in Figure 1, where the true parameters used are $\theta = (0.1, 0.01)$ and a prior uniform on the logarithm of each of the parameters over the interval $[10^{-3}, 10^3]$ was used. Here we show marginal posterior distributions generated using uniform weights, weights scaled with the standard deviation of each summary statistic, and adaptively chosen weights via the method outlined in Algorithm 2. We observe similar performance in identification of the decay parameter k using uniform weights, scaled weights and the adaptive choice of weights. Scaling the summary statistics with their standard deviation results in a posterior that does not provide much information over the prior for σ , since all the summary statistics are assumed to be equally informative which is not the case here. Note that only one summary statistic provides information about the random variable Z , whereas the other $n+1$ time points provide information about the decay of species A . However, the weights chosen via the search process outlined in Algorithm 2 give rise to a posterior that outperform the

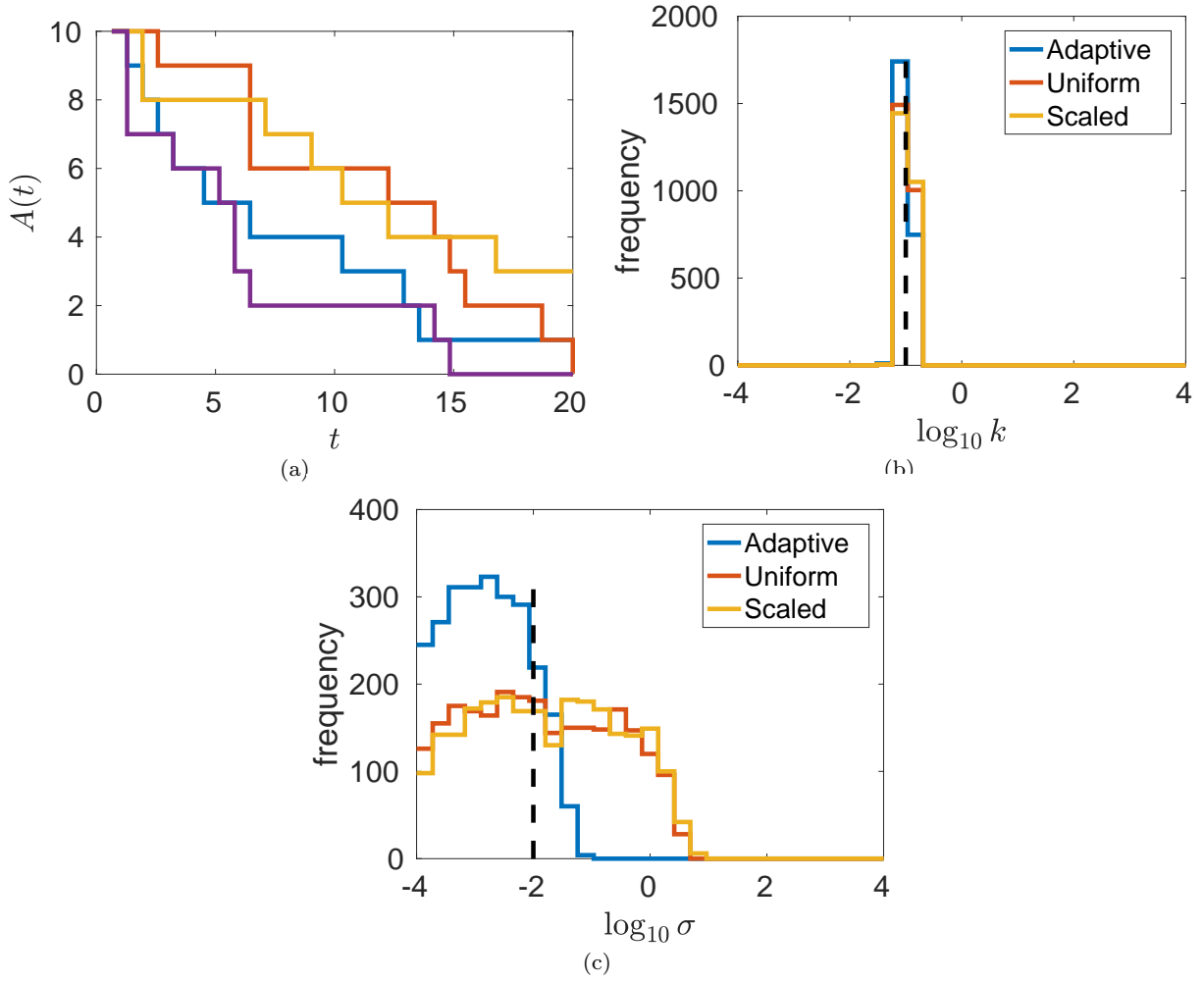


Fig. 1: Posteriors for parameters k and σ in the death process test problem for different weights in the ABC distance function. ABC-SMC was used to provide estimates of the posterior, with five generations and $N = 50,000$ simulations at each generation with the posterior constructed from the closest 5% of the simulations ($\alpha = 0.05$). (a) shows typical output from the model for the true parameters. The posteriors for k are given in (b) and for σ in (c).

posteriors generated using uniform weights and scaled weights for the second parameter σ , since only a single summary statistic provides relevant information for this parameter.

3.2 Birth-death process

Here we show similar results for a system with both first and second order reactions, representing both the production and degradation of a species A :



We assume known initial conditions of $A(0) = 1000$ and consider an observational time period of $[0, 40]$ with

$n = 32$ observations at regular intervals with no observational noise. For this system, we take the time series $S = [A(t_0), A(t_1), \dots, A(t_n)]$ as our summary statistics and seek to infer the two-dimensional parameter $\theta = (k_1, k_2)$.

Results of inference with ABC-SMC for this system are shown in Figure 2, with the methods for selecting the weights of the ABC distance function as for Figure 1. The true parameters used are $\theta = (0.01, 0.1)$ and we apply uniform priors on the logarithm of each of the parameters over the range $[10^{-4}, 10^0]$. The death parameter, k_1 , is clearly identified using both the uniform and adaptive choice of weights. Due to the initial condition, the system starts in a regime dominated by death events so this parameter is easier to identify.

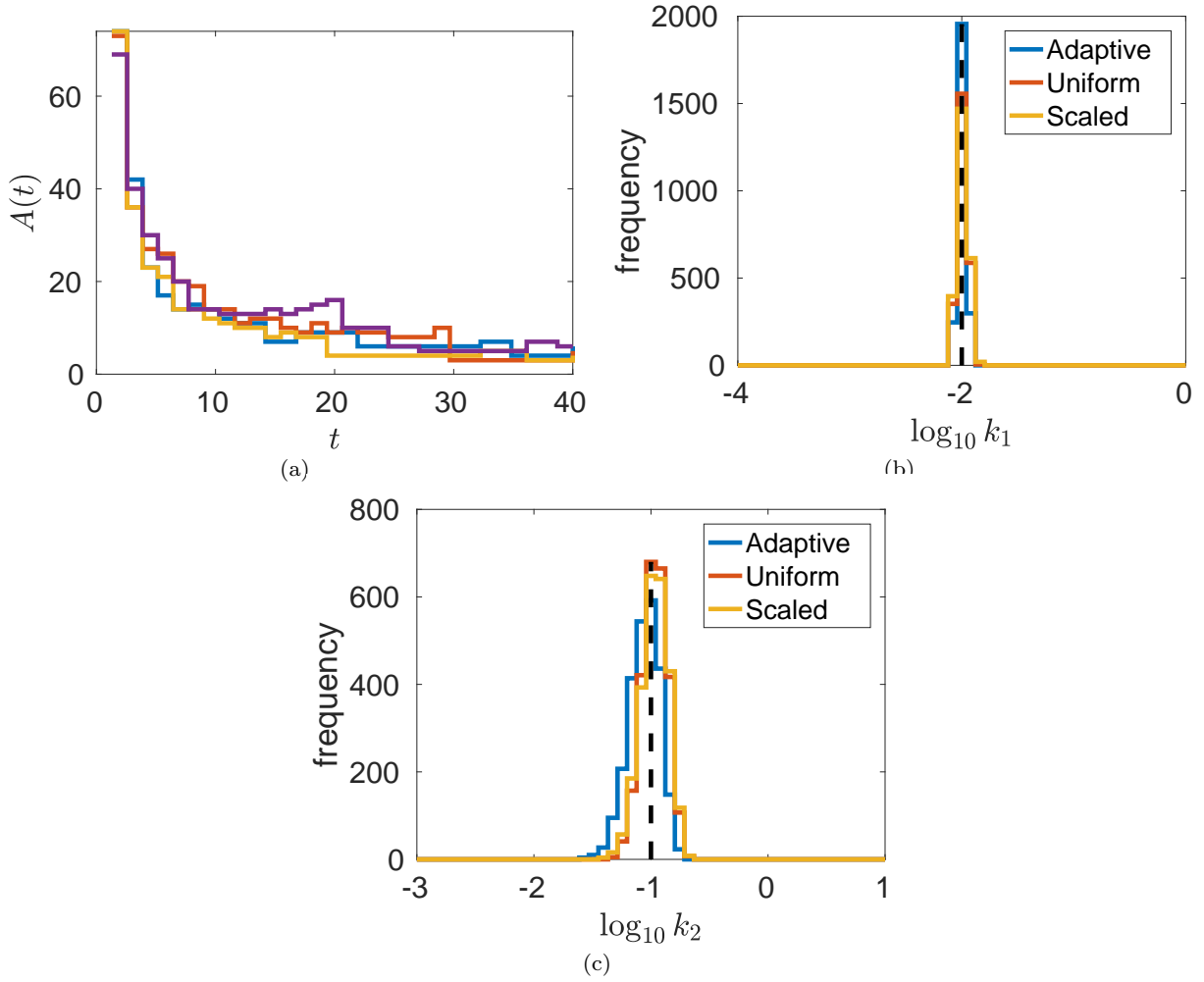


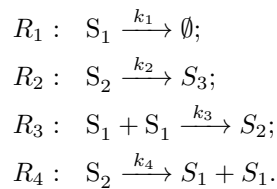
Fig. 2: Posteriors for parameters k_1 and k_2 in the birth-death process for different weights in the ABC distance function. ABC-SMC was used with five generations and $N = 50,000$ simulations at each generation with the posterior constructed from the closest 5% of the simulations ($\alpha = 0.05$). Note that (a) shows typical output from the model for the true parameters. Posterior marginal distributions for parameters k_1, k_2 are shown in (b) and (c), respectively.

The approximate posteriors for the birth parameter, k_2 , are much wider. They contain the true parameter value within their support and the adaptive method is able to exclude the largest area of parameter space.

3.3 Dimerization system

To examine a system with multiple scales, we consider also a dimerization system, which undergoes a fast initial transient followed by slower subsequent dynamics [19]. The dimerization system consists of the following

reactions:



We take initial conditions $S_1(0) = 10^5$, $S_2(0) = 0$, $S_3(0) = 0$ and consider an observational time period of $[0, 100]$ with $n = 32$ geometrically spaced observations (to capture the multiple timescales present), without observational noise. For the dimerization system, we take the time series $S = [S_1(t_0), \dots, S_1(t_n), S_2(t_0), \dots, S_2(t_n), S_3(t_0), \dots, S_3(t_n)]$ as our summary statistic and infer the four-dimensional parameter $\theta = (k_1, k_2, k_3, k_4)$.

We note that for a choice of parameter $\theta^* = (1, 0.04, 0.002, 0.5)$, and the given initial conditions, we obtain a fast decay of species S_1 and accumulation of species S_2 , followed by a slower decay of S_2 and accumulation of S_3 (see Figure 3(a)).

The results of parameter inference for this system can be seen in Figure 3. The true parameters used are $\theta = (1, 0.04, 0.002, 0.5)$, and we apply a prior uniform on the logarithm of the parameters over the intervals $[10^{-2}, 10^2]$, $[10^{-3}, 10^1]$, $[10^{-5}, 10^{-1}]$, $[10^{-3}, 10^1]$, respectively, for each of the parameters. Parameters k_1 and k_2 are clearly identified by the adaptive choice of weights. The fast transient behaviour initially involves reactions at rate k_1 , while k_2 corresponds to the longer timescale accumulation of species S_3 . Parameters k_3 and k_4 are harder to identify with broader resulting posteriors, but again the adaptive algorithm does a better job at excluding regions of search space than a uniform choice of weights, or a scaling with the standard deviation. Scaling by the standard deviation is a poor choice here because for some of the time points, particularly in the fast initial transient region, there is no variation between the synthetic datasets.

3.4 Simple spatial model

Spatial models produce very high dimensional data, containing information about dynamics in both space and time. Here, we consider a simple spatial model in one dimension to describe the spreading of particles by diffusion without volume exclusion. We divide our spatial domain $X \in [-1, 1]$ into m boxes or voxels, and label the numbers of particles in voxels $1, \dots, m$ as S_1, \dots, S_m , respectively. Particles can jump between neighbouring voxels at rate $\theta = D/h^2$, where D is the macroscopic diffusion constant and h is the width of the voxel. We assume zero flux conditions at $X = \pm 1$ and take $m = 8$, so that $h = 1/4$. As an initial condition, we place 10 particles in each of the $m/2$ voxels on the left-hand side of the domain where $x < 0$, and allow the system to evolve over the time interval $[0, 20]$. We observe the system at $n = 8$ equally spaced time points, and take as our summary statistic the time series for each voxel, $S = [S_1(t_0), \dots, S_1(t_n), S_2(t_0), \dots, S_2(t_n), \dots, S_m(t_0), \dots, S_m(t_n)]$, where $S_i(t_j)$ is the number of particles in voxel i at time point t_j . Using synthetic data simulated with $\theta = 0.1$, we attempt to recover the jump rate θ . The results of parameter inference for this problem are shown in Figure 4, where we have used a prior uniform on $\log_{10}(\theta)$ over the interval $[10^{-4}, 10^0]$. We successfully obtain an informative unbiased posterior for θ using the adaptive choice of weights, with a

notable improvement in comparison to the other methods for selecting the weights.

3.5 Computational overhead

If our proposed approach of adapting the weights of each of the summary statistics is to be used in practice, we must ensure that the increases in the quality of the resulting posterior justify the computational overhead required for the search process. Otherwise, it would be preferable simply to generate the posterior using ABC-SMC with more samples. Therefore we are interested in what the computational overhead of the search process is, and how to limit the cost of the search in higher dimensions. We note that the computational cost of the search process scales linearly with the number of samples simulated. This is because in comparing the distance of the approximate posterior from the prior, the approximate posterior distribution is discretized using a histogram at each search step.

Using the dimerization test problem, as described in Section 3.3, we ran Algorithm 2 with $N_1 = 5,000$, $\alpha_1 = 5\%$. To compare this to ABC-SMC with uniform weights, we performed parameter inference with uniform weights using both $N_1 = 5,000$, $\alpha_1 = 5\%$ and $N_2 = 5,600$, $\alpha_2 = 4.46\%$. The value of N_2 was chosen such that an equal length of computation time was spent in the search steps to find the summary statistic weights in Algorithm 2, as was spent in generating extra samples in ABC-SMC with uniform weights. A corresponding lower value of α was chosen so that the number of particles in the parameter sample was equivalent.

In this case, adaptively choosing weights using Algorithm 2 resulted in a significantly greater distance between the prior and posterior, and reduced the bias in the posterior compared to running ABC-SMC with more samples, as measured by the distance between the maximum posterior estimate and the true parameters. These results, which represent improvements in the posterior for the same computational cost, are shown in Table 1 and the same procedure was used for the other test problems.

3.6 Consistent weights

Ideally, our search process should find the global optimum weight vector, so that if Algorithm 2 is run multiple times the same weight vector is obtained. In practice, for the examples we have explored, the function to be optimized (distance between prior and posterior as a function of the distance weights) is very flat with

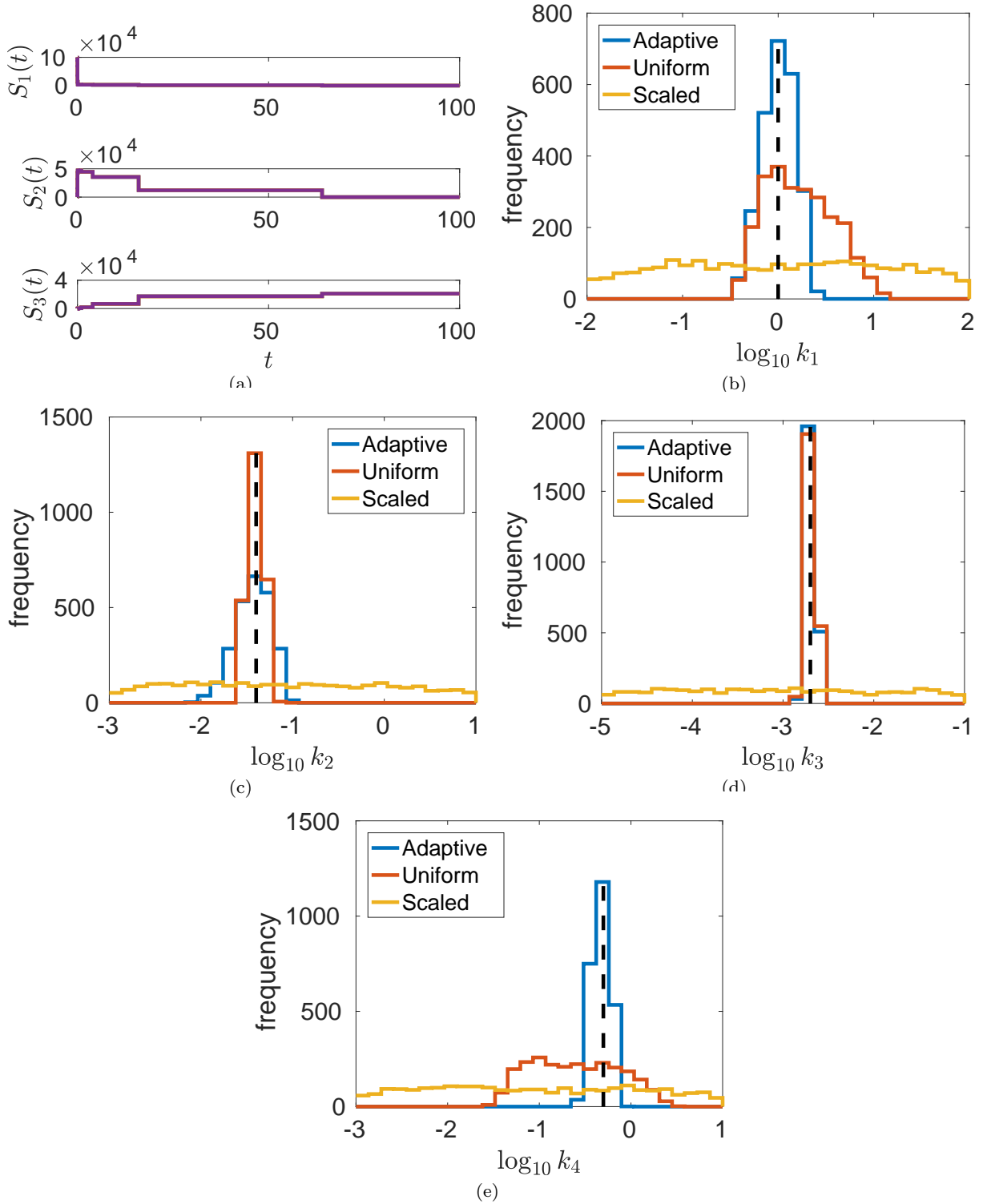


Fig. 3: Posteriors for parameters $\theta = (k_1, k_2, k_3, k_4)$ in the dimerization system for different weights in the ABC distance function. ABC-SMC was used with five generations and $N = 50,000$ simulations at each generation with the posterior constructed from the closest 5% of the simulations ($\alpha = 0.05$). (a) shows typical output from the model for the true parameters, for each species, S_i . Posterior marginal distributions for parameters k_1, k_2, k_3, k_4 are shown in (b) to (e).

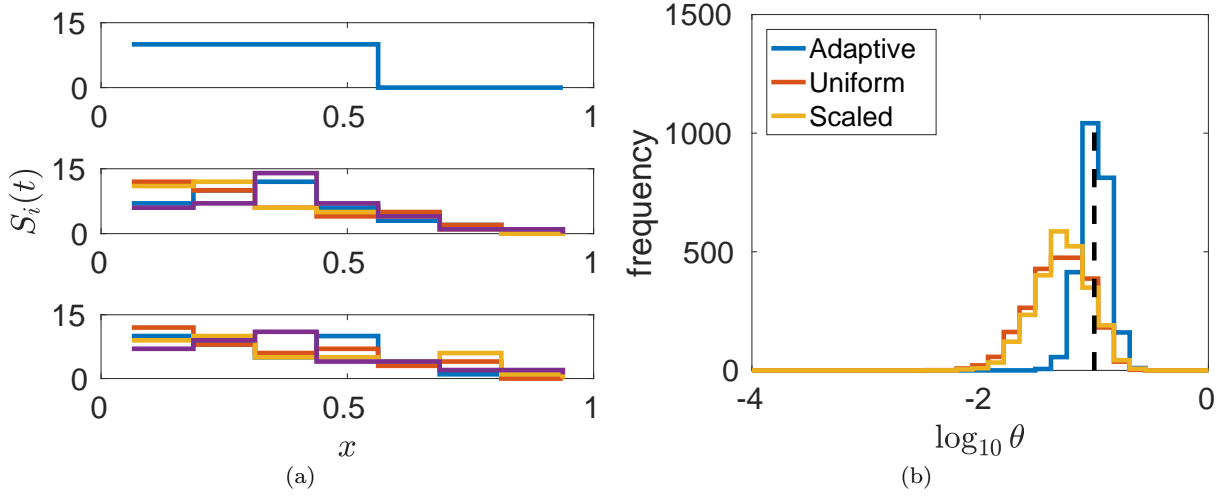


Fig. 4: Posteriors for parameter θ in the simple diffusion model for different weights in the ABC distance function. ABC-SMC was used for the inference with five generations and $N = 50,000$ simulations at each generation with the posterior constructed from the closest 5% of the simulations ($\alpha = 0.05$). (a) shows the spatial profile at three different time points ($t = 0, 10, 20$) and demonstrates the variability in the output for this spatial process across four realizations with the same parameter, $\theta = 0.1$. In (b), we compare the posteriors obtained for θ with different choices of weights.

Test problem	Hellinger distance between prior and posterior	Bias in posterior
Death process	0.926 / 0.929 / 0.929	0.69 / 1.35 / 0.29
Birth-death process	0.945 / 0.950 / 0.950	0.03 / 0.06 / 0.06
Dimerization	3.96 / 4.06 / 4.62	0.78 / 0.26 / 0.16
Diffusion	0.797 / 0.911 / 1.231	1.07 / 0.93 / 0.38

Table 1: Performance of Algorithm 2 compared with increasing the number of samples in ABC-SMC. Results are shown for each of the test problems in the form: ABC-SMC with N_1 and α_1 / ABC-SMC with N_2 and α_2 / Algorithm 2 with N_1 and α_1 . Highlighted in bold is the method with best performance according to each metric.

respect to some of the distance weights. As seen in Figure 5, where we explore how the chosen weights vary for Example 1, this makes it hard to consistently identify a global maximum and results in large variations in the distance weights. We can interpret this as the algorithm identifying the informative summary statistics and appropriately using the information from these, while allowing weights for other summary statistics to take a range of values without much effect on the resulting posterior. The weights found for different runs of the algorithm are highly correlated, however, as expected. To better compare the weights found by optimization across runs of the algorithm, we subtract the mean of the weights for each run of the Algorithm 2. This highlights the summary statistic z for the death process test problem as highly informative (see Figure 5 (c)), which agrees with our intuition, since only this summary statistic gives informative about the param-

eter σ , whereas any of the others can provide information about the decay parameter, k .

4 Discussion

In this work, we have presented a method for improving the quality of posteriors resulting from approximate inference using ABC-SMC by optimizing the weights of the ABC distance function, $d(x_1, x_2)$. By applying the methodology to several test problems, we have demonstrated that our novel, adaptive method allows effective combination of summary statistics. We see superior performance using our algorithm in comparison with naive choices of uniform weights or using the scale of the summary statistics. Further benefits of adapting the weights include removing the requirement for design and selection of summary statistics ‘by hand’.

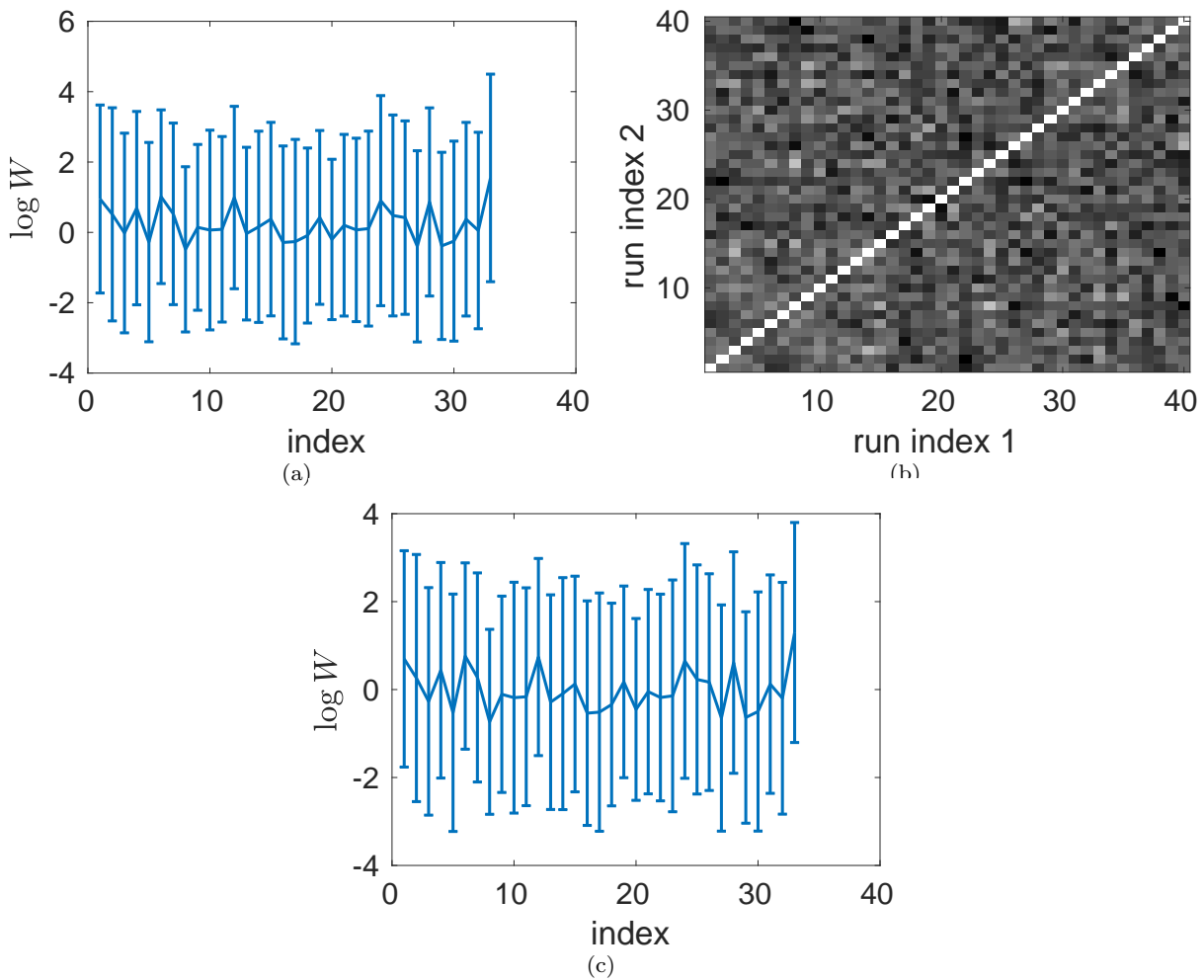


Fig. 5: The optimum distance weights found from the search procedure after 40 successive runs of Algorithm 2 on the death process test problem described in Section 3.1 with parameters as for Figure 1. In (a), we show the logarithm of the raw weights found by the search procedure, which can be compared with (c) showing the difference between each weight and mean of the weights for each run. In (b), we show the correlation between the weights across repeated runs of the ABC distance weight algorithm.

4.1 Comparison to dimensionality reduction methods

Adaptively choosing the summary statistic weights within the ABC distance function can be seen as achieving a similar goal to summary statistic dimension reduction techniques [8, 21]. These techniques either project high-dimensional summary statistics into a lower dimensional subspace, or select an optimal subset of summary statistics via some optimality criterion. In contrast, a similar effect is achieved here when the statistics are combined in the weighted Euclidean distance function, $d(x_1, x_2)$, by weighting summary statistics to take account of both their inherent scale, but also their relative contribution towards the posterior distribution. Uninformative summary statistics are automatically as-

signed a lower weighting, while more informative summary statistics are given high weights relative to their scale.

Previous subset selection methods have used criteria for approximate sufficiency of a subset of statistics to test whether adding a new statistic results in a change in the posterior above a certain threshold [17]; minimising an information criterion based on knn-entropy over all subsets of summary statistics [21]; and reducing loss of information by adding summary statistics until the KL divergence between the resulting posteriors is below a threshold [3]. All of these methods seek to choose a lower dimensional subset of a given list of summary statistics. Using this lower dimensional subset increases the acceptance rate for samples in ABC

Test problem	Hellinger distance between prior and posterior	Entropy	Bias in posterior
Death process	2.75 / 2.40 / 0.80	3.87 / 3.96 / 14.95	1.31 / 1.31 / 1.61
Birth-death process	1.60 / 1.09 / 1.27	10.93 / 14.56 / 11.49	1.54 / 2.24 / 1.73
Dimerization	2.72 / 2.20 / 2.31	33.14 / 39.70 / 34.8	2.00 / 1.07 / 0.18
Diffusion	1.63 / 0.81 / 1.72	2.48 / 3.69 / 2.37	0.03 / 0.10 / 0.03

Table 2: Comparison of the quality of the posteriors obtained using different methods to combine summary statistics. Results given as adaptive method/Barnes et al. [3]/Fearnhead and Prangle [10]. Bold text highlights the best performance on a metric for a test problem.

by avoiding the curse of dimensionality for the data. However, the results depend on the order in which the summary statistics (or subsets) are analysed.

A popular method, implemented in packages such as abctools [22], is the semi-automatic ABC approach of Fearnhead and Prangle [10]. This approach uses a projection method to find informative linear combinations of statistics by fitting a regression for each parameter in the model. The result is a reduction from the original high-dimensional set of summary statistics to new lower dimensional set of summary statistics with the same dimension as the parameter space. Improved results are seen by using a pilot run of ABC to choose a subset of parameter space as a training region for the regression. Further improvements are obtained by extending the vector of summary statistics by concatenating with a non-linear transformation of the same summary statistics, $S = (s, s^2, s^3, s^4)$, where s is a given vector of summary statistics and the superscripts indicating raising these to the given power. This method uses contributions from all of the summary statistics and should optimize the mean quadratic loss.

We tested our adaptive weight selection algorithm against the semi-automatic ABC method [10], and the subset selection method of Barnes et al. [3] based on an approximate sufficiency criterion. In general, for the test problems considered, as described in Section 3, our method outperforms the competing methods, as shown by the metrics in Table 2. We note that a lower value of the entropy in this case corresponds to a tighter, more informative posterior. A larger value of the Hellinger distance indicates a greater distance between prior and posterior. The bias gives the distance between the maximum posterior estimate and the true parameter value. In implementing these methods, we have used only ABC rejection sampling, equivalent to a single generation of ABC-SMC, to compare the methods. In practice, these results mean that our method outlined in Algorithm 2 for adaptively choosing the weights of summary statistics produces a more informative posterior than competing methods based on dimensionality reduction of summary statistics.

4.2 Further work

Our method for automatically adapting the weights of the ABC distance function could be combined with other methods for dimensionality reduction of summary statistics to further improve the quality of posteriors produced with ABC for given computational effort. A particular area to consider would be how best to combine optimization of the distance weights for ABC and dimensionality reduction of the summary statistics. These are related approaches that can work well together. One approach that could be explored, for example, is enforcing some sparsity of the weights during the search step of the weights optimization. By setting some weights to be explicitly zero, we exclude the corresponding summary statistics, effectively reducing the dimensionality of our summary statistics. Further investigations could explore how best to sample sparse subsets of weights in high dimensions.

4.3 Conclusion

In summary, we propose a computationally efficient search procedure to identify a set of optimum weights to allow us to combine summary statistics within the ABC distance function in such a way that the gain in information in the posterior over the prior is maximized.

Acknowledgements This work was supported by funding from the Engineering and Physical Sciences Research Council (EPSRC) (grant no. EP/G03706X/1).

A Table of hyperparameters

References

1. Andrieu C, Roberts GO (2009) The pseudo-marginal approach for efficient Monte Carlo computations. *The Annals of Statistics* 37(2):697–725
2. Andrieu C, Doucet A, Holenstein R (2010) Particle Markov chain Monte Carlo methods. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 72(3):269–342

Test problem	n	A(0)	T	θ^*	N	α	Repeats	Search steps	λ	Proposal s.d.	Prior interval
Death process	32	10	20	(0.1, 0.01)	$5 * 10^4$	0.05	5	$5 * 10^4$	$2 * 10^{-4}$	0.25	$[10^{-4}, 10^4]$
Birth-death process	32	10^3	40	(0.1, 0.1)	$5 * 10^4$	0.05	5	$5 * 10^4$	$2 * 10^{-4}$	0.25	$[10^{-4}, 10^0]$
Dimerization	32	$(10^5, 0, 0)$	100	(1, 0.04, 0.002, 0.5)	$5 * 10^4$	0.05	1	$5 * 10^3$	$2 * 10^{-4}$	0.25	$[10^{-2}, 10^2]$, $[10^{-3}, 10^1]$, $[10^{-5}, 10^{-1}]$, $[10^{-3}, 10^1]$
Diffusion	8	$10 * \mathbb{1}_{x < 0}$	20	0.1	$5 * 10^4$	0.05	5	$5 * 10^3$	$2 * 10^{-4}$	0.25	$[10^{-4}, 10^0]$

Table 3: Summary of hyperparameters used in simulations.

3. Barnes C, Filippi S, Stumpf M, Thorne T (2012) Considerate approaches to constructing summary statistics for ABC model selection. *Statistics and Computing* 22(6):1181–1197
4. Barthelmé S, Chopin N (2014) Expectation propagation for likelihood-free inference. *Journal of the American Statistical Association* 109(505):315–333
5. Beaumont MA (2010) Approximate Bayesian computation in evolution and ecology. *Annual Review of Ecology, Evolution and Systematics* 41:379–405
6. Beaumont MA, Zhang W, Balding DJ (2002) Approximate Bayesian computation in population genetics. *Genetics* 162(4):2025–2035
7. Bishop CM (2006) *Pattern Recognition and Machine Learning*. Springer
8. Blum MG, Nunes MA, Prangle D, Sisson SA (2013) A comparative review of dimension reduction methods in approximate Bayesian computation. *Statistical Science* 28(2):189–208
9. Del Moral P, Doucet A, Jasra A (2006) Sequential Monte Carlo samplers. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 68(3):411–436
10. Fearnhead P, Prangle D (2012) Constructing summary statistics for approximate Bayesian computation: semi-automatic approximate Bayesian computation. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 74(3):419–474
11. Gillespie DT (1977) Exact stochastic simulation of coupled chemical reactions. *Journal of Physical Chemistry* 81(25):2340–2361
12. Golightly A, Wilkinson DJ (2011) Bayesian parameter inference for stochastic biochemical network models using particle Markov chain Monte Carlo. *Interface Focus* 1(6):807
13. Gourieroux C, Monfort A, Renault E (1993) Indirect inference. *Journal of Applied Econometrics* 8(S1):85–118
14. Hartig F, Calabrese JM, Reineking B, Wiegand T, Huth A (2011) Statistical inference for stochastic simulation models—theory and application. *Ecology Letters* 14(8):816–827
15. Hines KE (2015) A primer on Bayesian inference for biophysical systems. *Biophysical Journal* 108(9):2103–2113
16. Jones P, Sim A, Taylor H, Bugeon L, Dallman M, Pereira B, Stumpf M, Liepe J (2015) Inference of random walk models to describe leukocyte migration. *Physical Biology* 12(6):66,001–66,012
17. Joyce P, Marjoram P (2008) Approximately sufficient statistics and Bayesian computation. *Statistical Applications in Genetics and Molecular Biology* 7(1)
18. Jung H, Marjoram P (2011) Choice of summary statistic weights in approximate Bayesian computation. *Statistical Applications in Genetics and Molecular Biology* 10(1):1–23
19. Lester C, Yates CA, Giles MB, Baker RE (2015) An adaptive multi-level simulation algorithm for stochastic biological systems. *The Journal of Chemical Physics* 142(2):024,113
20. McKinley T, Cook AR, Deardon R (2009) Inference in epidemic models without likelihoods. *The International Journal of Biostatistics* 5(1):1–40
21. Nunes MA, Balding DJ (2010) On optimal selection of summary statistics for approximate Bayesian computation. *Statistical Applications in Genetics and Molecular Biology* 9(1):1–16
22. Nunes MA, Prangle D (2015) abctools: an R package for tuning Approximate Bayesian Computation analyses. *The R Journal* 7(2):189–205
23. Owen J, Wilkinson DJ, Gillespie CS (2015) Scalable inference for Markov processes with intractable likelihoods. *Statistics and Computing* 25(1):145–156
24. Prangle D (2015) Adapting the ABC distance function. *arXiv preprint arXiv:150700874*
25. Price LF, Drovandi CC, Lee A, Nott DJ (2016) Bayesian synthetic likelihood, preprint available at <http://eprints.qut.edu.au/92795/>
26. Pritchard JK, Seielstad MT, Perez-Lezaun A, Feldman MW (1999) Population growth of human Y chromosomes: a study of Y chromosome microsatellites. *Molecular Biology and Evolution* 16(12):1791–1798
27. Sisson SA, Fan Y, Tanaka MM (2007) Sequential Monte Carlo without likelihoods. *Proceedings of the National Academy of Sciences* 104(6):1760–1765
28. Sunnåker M, Busetto AG, Numminen E, Corander J, Foll M, Dessimoz C (2013) Approximate Bayesian computation. *PLOS Computational Biology* 9(1):e1002,803
29. Toni T, Welch D, Strelkowa N, Ipsen A, Stumpf M (2009) Approximate Bayesian computation scheme for parameter inference and model selection in dynamical systems. *Journal of the Royal Society Interface* 6(31):187–202
30. Turner BM, Van Zandt T (2012) A tutorial on approximate Bayesian computation. *Journal of Mathematical Psychology* 56(2):69–85
31. Wilkinson DJ (2009) Stochastic modelling for quantitative description of heterogeneous biological systems. *Nature Reviews Genetics* 10(2):122–133
32. Wood SN (2010) Statistical inference for noisy nonlinear ecological dynamic systems. *Nature* 466(7310):1102–1104

